Peer Review File

Article Information: https://dx.doi.org/10.21037/tlcr-21-895

Reviewer A

Comment 1: Definitely the role of radiomics predicting STAS need further development and would hopefully help to improve daily practice clinical decisions. In this context I find this study very encouraging.
However, taking into account the methodology, did the authors include any clinical variables that could confound/alter the results of their model? What was the AUC of clinical characteristics alone? I think adding this data could help improve the value of this manuscript

Reply 1: Thank you very much for your kind and useful comment. The aim of the study was to develop a STAS prediction radiomics model that was independent from clinical and radiological variables. This is the main reason why we did not include any clinical variable in the model. Moreover, in our cohort the clinical characteristics don't allow a reliable prediction of STAS with the best model showing an AUC of 0,54. As suggested by the reviewer this data could improve the comprehension of the study and it was added in a dedicated table (Table 4).

Changes in the text: Table 4 has been added including AUC of the clinical characteristics

Reviewer B

Thank you very much for your supportive comments and suggestions. Please deal with the following answers aimed at improving the quality of submission.

Comment 1: The authors normalized all CT images to 1 mm pixel size and 2 mm slice thickness. What would be the effect of selected normalization on selected features in final model that was used for cancer prediction.

Reply 1: The need for a normalization of data is crucial with a heterogeneous dataset as widely reported in other radiomics studies. In the final model, we selected only those features that maintain their predictive power after normalization process, as explained in the material and method section. However, the effect of normalization over any single features was not an object of the study and it could be an interesting topic for future investigations. Surely, as the reviewer pointed out, it represents a limitation of our study. A sentence has been added in the discussion section to highlight this point of weakness.

Changes in the text: Please refer to the discussion section lines 340-341 for the requested changes.

Comment 2: Why did authors use preprocessing using gray levels of bin size of 25; is there specific reasons for this selection?

Reply 2: As it was for volume distribution, gray levels have also to be normalized to remove the dependency of the selected feature from gray level distribution. Moreover, a reduced color deepness (i.e. the number of

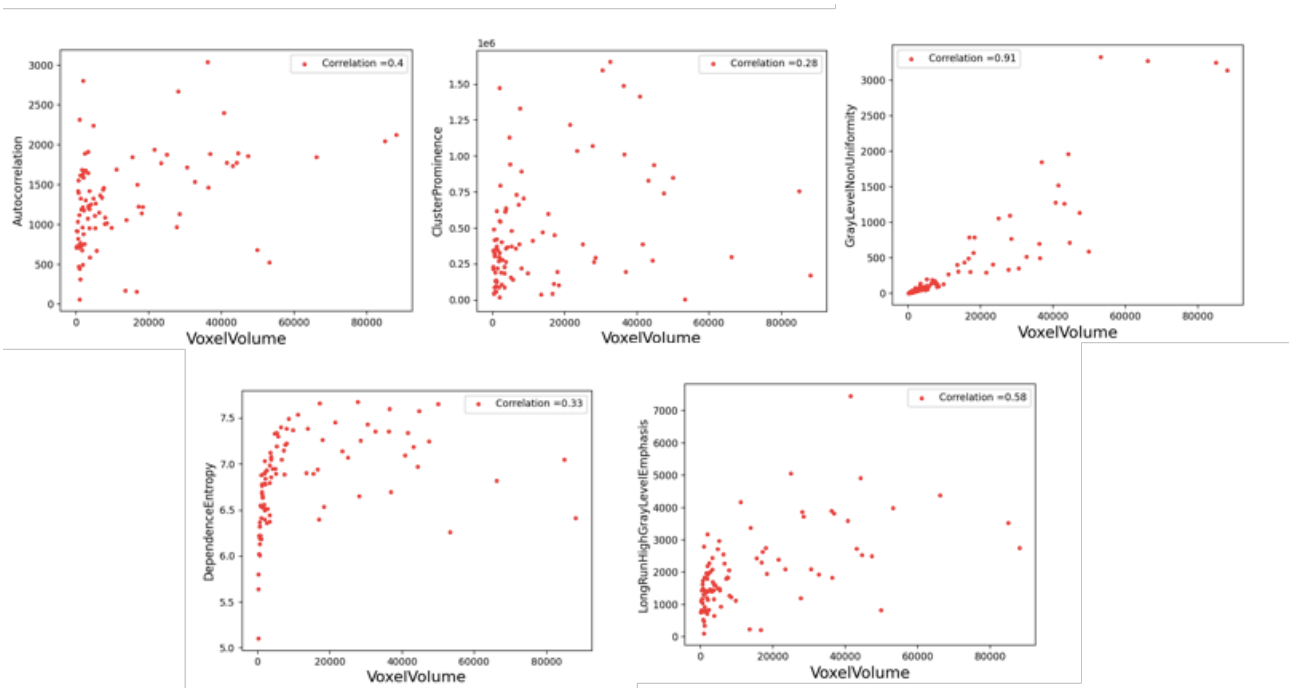
gray levels) improves the stability of the feature extraction. In our opinion, 25 bin width is a good balance between stability of the feature extraction and complexity of the image.

Changes in the text: Please refer to the Discussion section lines 306 - 308 for the requested changes.

Comment 3: Also, some features are dependent on the size of the tumor volume, is this issue of volume dependency addressed in this manuscript, the authors might need to discuss this important aspect of radiomics features.

Reply 3: Tumor volume is one of the variables that we considered for the prediction model as well as other "shape features" (shape, surface area, sphericity, maximum 3D diameter etc.). In our final model, only *GreyLevelNonUniformity* is strongly correlated with volume with a correlation index of 0.9 (please refer to figures below for graphical insight). This could explain why no "shape features" entered the final model since we put a cut on correlation. Probably *GreyLevelNonUniformity* contains more information than "shape features". We add a short comment on this topic in the discussion section.

Changes in the text: Please refer to Discussion section lines 318-322 for the dedicated changes.



Comment 4: What kind of reconstruction kernel is used in data acquisition? This need to be specified and it's affect needs to be discussed?

Reply 4: Thank you for the valuable question. In our dataset, the images had a high variety in the acquisition protocol with 19 different reconstruction kernels depending on the hospital center of origin. When possible, the images were acquired with a sharp (B60) or very sharp (70) reconstruction kernel. One of the objectives of the study was to create a reliable model using images that were non-homogeneous for acquisition protocol (coming from different sources). Indeed, the reconstruction kernel heterogeneity reflects the daily clinical practice in which patients are referred from different remote hospitals.

Changes in the text: Please find in "Imaging acquisition and segmentation" subsection, lines 150 – 151, the specified method of reconstruction kernel selection.

Reviewer C

Thank you for your encouraging comment. Please find below the changes made on the manuscript following your useful suggestions.

Comment 1: Since the intent of the study is to have a radiological/radiomics algorithm help preoperatively determining sublobar vs. lobar resections in stage I, small lung adenocarcinomas, it would be ideal to include stage I tumors only in the study cohorts or at least in the validation cohort.

Reply 1: The primary aim of the study was to test a radiomics based prediction model on a "real life" cohort of patients. Just for this purpose, the population is not homogeneous for quality of imaging and tumor characteristics. The core aim of the study excludes the possibility to arbitrarily subset patients. For this reason, we choose not to select patients for tumor stage but to enroll them consecutively. Ultimately, the future application of this prediction model will be centered on stage I lung cancer patients, where the option for a sublobar or lobar resection is still debated and lacks evidence. Further investigations on stage I lung cancer are needed and it will probably be the object of future studies. However, we believe that a study including not only stage I lung cancer is crucial to understand the behavior of the radiomics features in this context.

Anyway, we think that an aside exploratory analysis on stage I subset of patients could be worthy as suggested by the reviewer. The analysis of our models on the 31 stage I patients in the validation cohort showed similar results (best accuracy 0.74 – mixed model).

Changes in the text: The results of exploratory analysis on Stage I patients are reported in the Results section. Please refer to lines 245 - 247 in the Results section.

Comment 2: The authors acknowledged that the models are far from clinical application in the discussion, but it should formally be noted that specificity of 52.6% - 63.4% could lead to unnecessary lobectomy in up to 40% of patients, although specificity data specific for stage I tumors were not provided.

Reply 2: We understand your remark and a sentence was added formally specifying the aforementioned consequences. As pointed out by the reviewer, more truthful data will be provided by a specific-stage I lung cancer study. However, to consider lobectomy unnecessary for a stage I lung cancer is still controversial to date.

Changes in the text: Please refer to discussion section lines 327-330 for the suggested changes.

Comment 3: To accurately assess the presence or absence of STAS in a confident manner, ample benign parenchyma adjacent to the tumor needs to be assessed to overcome a heterogeneity issue. Assessing only one histology slide with benign, adjacent lung parenchyma may not be sufficient in this context depending on the

size of tumor. I would suggest that those with benign, adjacent lung parenchyma surrounding less than 1/3 of the tumor circumference available need to be excluded from the study.

Reply 3: We completely agree with the Reviewer's comments. Indeed, all tumors included in the study show a circumferential rim of normal lung parenchyma. To be more accurate we excluded from the study those resections where the benign surrounding parenchyma were less than 1 cm in all sections (e.g. resection margins < 1 cm). STAS was evaluated circumferentially and in multiple histological slices and not only in 1 slide as could be interpreted reading the unrevised material and methods section. We have modified the text to clarify this issue.

Changes in the text: Please refer to "histological evaluation" subsection for the aforementioned changes.

Comment 4: It has already been shown that the diagnosis of STAS is known to be subjective. The two pathologists reviewed each specimen individually? If so, what was their concordance on STAS? If there were discrepant interpretations between them, how was the final interpretation made?

Reply 4: In our institution the presence of STAS is routinely assessed in lung cancer. The two pathologists, with more than 10 years' expertise in lung cancer pathology, separately reviewed each slide of surgical specimens. There were no major discrepancies in the evaluation of presence and extension of STAS. Potential artifacts were excluded. Selected cases were discussed and any minor discrepancies between the two pathologists were resolved through discussion until consensus was reached.

Changes in the text: Please refer to "histological evaluation" subsection lines 141 - 142 for the aforementioned changes.

Comment 5: No significant difference in the histologic subtype and tumor stage (T, N, overall) was found between STAS positive and STAS negative patients. It does not seem in accordance with the prior reports. The discrepancy may be attributed in part to the small size of the cohorts of this study.

Reply 5: We agree with the reviewer. We also attribute this result to the small number of patients that is certainly one of the major limitations of the study as declared already.

Changes in the text: No changes were made accordingly

Comment 6: Incomplete surgical resection (R1) was considered as an exclusion criterion in the validation cohort, but not in the training cohort. What is a reason for the discrepancy?

Reply 6: Incomplete surgical resection was considered an exclusion criteria also for the training cohort. Actually we declared a complete surgical resection (R0) in the inclusion criteria. The verbal discrepancy has been fixed for a better comprehension.

Changes in the text: Refer to subsection 2.1, line 116 for the requested changes.

Comment 7: Page 5, lines 30-31: Moreover, when two features showed a high correlation (Pearson's correlation > 0.7), the one with the lowest p-value was excluded. – I don't understand why the lowest p-value, not the highest p-value, was excluded.

Reply 7: This is a transcription error. When two features showed a high correlation index, we exclude the one with the weakest p-value (e.g. the highest p-value).

Changes in the text: The sentence was changed accordingly (see subsection 2.5 line 185).

Comment 8: The process of incorporating both radiomics and radiologic characteristics to build the mixed model should be described in detail.

Reply 8: For the mixed model, all radiological and radiomics extracted features together entered in the mixed classifier. Features were firstly tested for significance and then for correlation independently. Finally, the most representative 5 features were used for training and subsequently testing the model. Our selection pipeline automatically selected 3 radiomics and 2 radiological features (e.g. mixed model) without knowing the origin of the feature.

Changes in the text: Please find in the Results section lines 236 - 238 a sentence describing the incorporation of radiomics and radiological features in the mixed model.

Comment 9: A table to show the results of multivariate analysis in radiologic characteristics would be helpful.

Reply 9: A table resuming the results of multivariate analysis of the different predictors was added.

Changes in the text: Table 4 was added as requested.

Comment 10: Table 3: what is Autocorrelation?

Reply 10: It is the radiomics feature name. It is defined as follows: "Autocorrelation is a measure of the magnitude of the fineness and coarseness of texture"

(https://pyradiomics.readthedocs.io/en/latest/features.html)

Changes in the text: No changes are needed

Comment 11: Tumor stage was based on AJCC 7th or 8th?

Reply 11: All patients were staged according to the AJCC TNM classification 8th edition. As suggested this has been specified in the revised manuscript.

Changes in the text: Please refer to subsection 3.1 lines 214-215 for the aforementioned changes.